1. Chapter 9 - Simple Linear Regression
    1. Introduction -- Linear Regression and Correlation -- MtRoyal - Version2016RevA
    2. Linear Equations -- Linear Regression and Correlation -- MtRoyal - Version2016RevA
    3. Scatter Plots -- Linear Regression and Correlation -- MtRoyal - Version2016RevA

Introduction -- Linear Regression and Correlation -- MtRoyal - Version2016RevA
class="introduction"

Linear regression and correlation can help you determine if an auto mechanic’s salary is related to his work experience. (credit: Joshua Rothhaas)

Note:
Chapter Objectives
By the end of this chapter, the student should be able to:

- Discuss basic ideas of linear regression and correlation.
- Create and interpret a line of best fit.
- Calculate and interpret the correlation coefficient.
- Calculate and interpret outliers.

Professionals often want to know how two or more numeric variables are related. For example, is there a relationship between the grade on the second math exam a student takes and the grade on the final exam? If there is a relationship, what is the relationship and how strong is it?

In another example, your income may be determined by your education, your profession, your years of experience, and your ability. The amount you pay a repair person for labor is often determined by an initial amount plus an hourly fee.

The type of data described in the examples is **bivariate** data — "bi" for two variables. In reality, statisticians use **multivariate** data, meaning many variables.

In this chapter, you will be studying the simplest form of regression, "linear regression" with one independent variable (x). This involves data that fits a line in two dimensions. You will also study correlation which measures how strong the relationship is.

Linear Equations -- Linear Regression and Correlation -- MtRoyal - Version2016RevA

Linear regression for two variables is based on a linear equation with one independent variable. The equation has the form:
Equation:

$$y = a + \mathrm{bx}$$

where *a* and *b* are constant numbers.

The variable ***x* is the independent variable, and *y* is the dependent variable.** Typically, you choose a value to substitute for the independent variable and then solve for the dependent variable.

Example:
The following examples are linear equations.
Equation:

$$y = 3 + 2\mathrm{x}$$

Equation:

$$y = -0.01 + 1.2\mathrm{x}$$

Note:
Try It
Exercise:

Problem: Is the following an example of a linear equation?

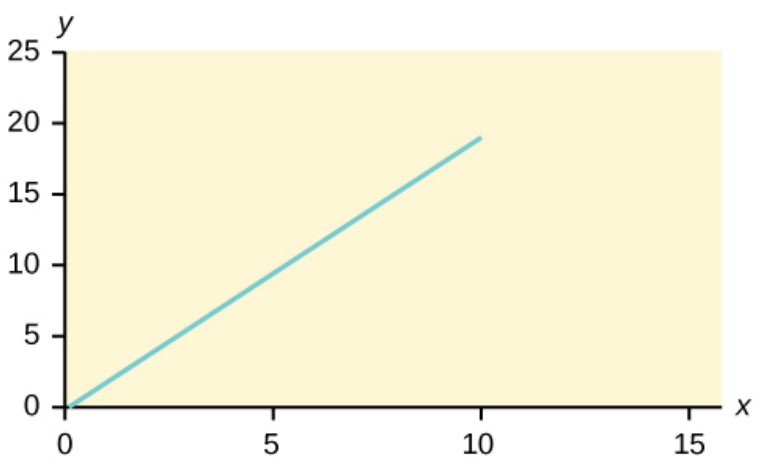
$y = -0.125 - 3.5x$

Solution:

yes

The graph of a linear equation of the form $y = a + bx$ is a **straight line**. Any line that is not vertical can be described by this equation.

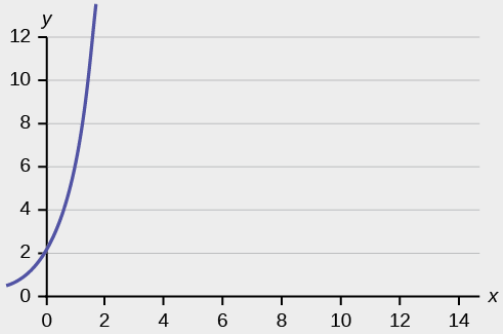
Example:
Graph the equation $y = -1 + 2x$.



Note:
Try It
Exercise:

Problem:

Is the following an example of a linear equation? Why or why not?



Solution:

No, the graph is not a straight line; therefore, it is not a linear equation.

Example:
Aaron's Word Processing Service (AWPS) does word processing. The rate for services is $32 per hour plus a $31.50 one-time charge. The total cost to a customer depends on the number of hours it takes to complete the job.
Exercise:

Problem:

Find the equation that expresses the **total cost** in terms of the **number of hours** required to complete the job.

Solution:

Let x = the number of hours it takes to get the job done.
Let y = the total cost to the customer.

The $31.50 is a fixed cost. If it takes x hours to complete the job, then (32)(x) is the cost of the word processing only. The total cost is: $y = 31.50 + 32x$

Note:
Try It
Exercise:

Problem:

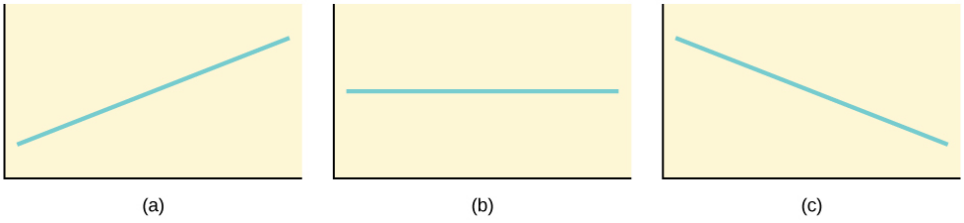
Emma’s Extreme Sports hires hang-gliding instructors and pays them a fee of $50 per class as well as $20 per student in the class. The total cost Emma pays depends on the number of students in a class. Find the equation that expresses the total cost in terms of the number of students in a class.

Solution:

$y = 50 + 20x$

Slope and *Y*-Intercept of a Linear Equation

For the linear equation $y = a + bx$, b = slope and $a = y$-intercept. From algebra recall that the slope is a number that describes the steepness of a line, and the y-intercept is the y coordinate of the point $(0, a)$ where the line crosses the y-axis.



Three possible graphs of $y = a + bx$. (a) If $b > 0$, the line slopes upward to the right. (b) If $b = 0$, the line is horizontal. (c) If $b < 0$, the line slopes downward to the right.

Example:
Svetlana tutors to make extra money for college. For each tutoring session, she charges a one-time fee of $25 plus $15 per hour of tutoring. A linear equation that expresses the total amount of money Svetlana earns for each session she tutors is $y = 25 + 15x$.

Exercise:

Problem:

What are the independent and dependent variables? What is the y-intercept and what is the slope? Interpret them using complete sentences.

Solution:

The independent variable (x) is the number of hours Svetlana tutors each session. The dependent variable (y) is the amount, in dollars, Svetlana earns for each session.

The y-intercept is 25 ($a = 25$). At the start of the tutoring session, Svetlana charges a one-time fee of $25 (this is when $x = 0$). The slope is 15 ($b = 15$). For each session, Svetlana earns $15 for each hour she tutors.

Note:
Try It

Exercise:

Problem:

Ethan repairs household appliances like dishwashers and refrigerators. For each visit, he charges $25 plus $20 per hour of work. A linear equation that expresses the total amount of money Ethan earns per visit is $y = 25 + 20x$.

What are the independent and dependent variables? What is the y-intercept and what is the slope? Interpret them using complete sentences.

Solution:

The independent variable (x) is the number of hours Ethan works each visit. The dependent variable (y) is the amount, in dollars, Ethan earns for each visit.

The y-intercept is 25 ($a = 25$). At the start of a visit, Ethan charges a one-time fee of $25 (this is when $x = 0$). The slope is 20 ($b = 20$). For each visit, Ethan earns $20 for each hour he works.

References

Data from the Centers for Disease Control and Prevention.

Data from the National Center for HIV, STD, and TB Prevention.

Chapter Review

The most basic type of association is a linear association. This type of relationship can be defined algebraically by the equations used, numerically with actual or predicted data values, or graphically from a plotted curve. (Lines are classified as straight curves.) Algebraically, a linear equation typically takes the form $\boldsymbol{y = mx + b}$, where $\boldsymbol{m}$ and $\boldsymbol{b}$ are constants, $\boldsymbol{x}$ is the independent variable, $\boldsymbol{y}$ is the dependent variable. In a statistical context, a

linear equation is written in the form $\boldsymbol{y = a + bx}$, where $\boldsymbol{a}$ and $\boldsymbol{b}$ are the constants. This form is used to help readers distinguish the statistical context from the algebraic context. In the equation $y = a + bx$, the constant b that multiplies the $\boldsymbol{x}$ variable (b is called a coefficient) is called as the **slope**. The slope describes the rate of change between the independent and dependent variables; in other words, the rate of change describes the change that occurs in the dependent variable as the independent variable is changed. In the equation $y = a + bx$, the constant a is called as the y-intercept. Graphically, the y-intercept is the y coordinate of the point where the graph of the line crosses the y axis. At this point $x = 0$.

The **slope of a line** is a value that describes the rate of change between the independent and dependent variables. The **slope** tells us how the dependent variable (y) changes for every one unit increase in the independent (x) variable, on average. The ***y*-intercept** is used to describe the dependent variable when the independent variable equals zero. Graphically, the slope is represented by three line types in elementary statistics.

Formula Review

$y = a + bx$ where a is the y-intercept and b is the slope. The variable x is the independent variable and y is the dependent variable.

Use the following information to answer the next three exercises. A vacation resort rents SCUBA equipment to certified divers. The resort charges an up-front fee of $25 and another fee of $12.50 an hour.

Exercise:

Problem: What are the dependent and independent variables?

Solution:

dependent variable: fee amount; independent variable: time

Exercise:

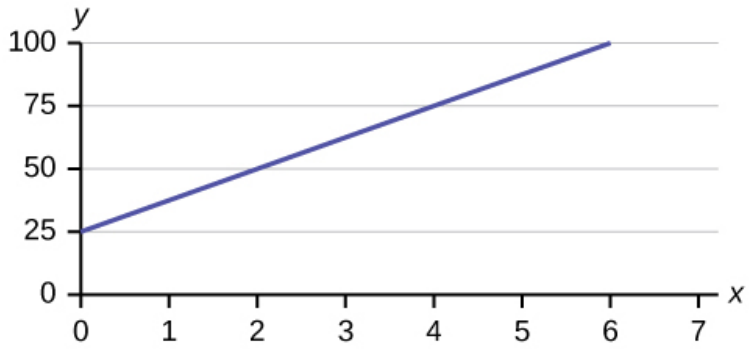
Problem:

Find the equation that expresses the total fee in terms of the number of hours the equipment is rented.

Exercise:

Problem: Graph the equation from [link].

Solution:



Use the following information to answer the next two exercises. A credit card company charges $10 when a payment is late, and $5 a day each day the payment remains unpaid.

Exercise:

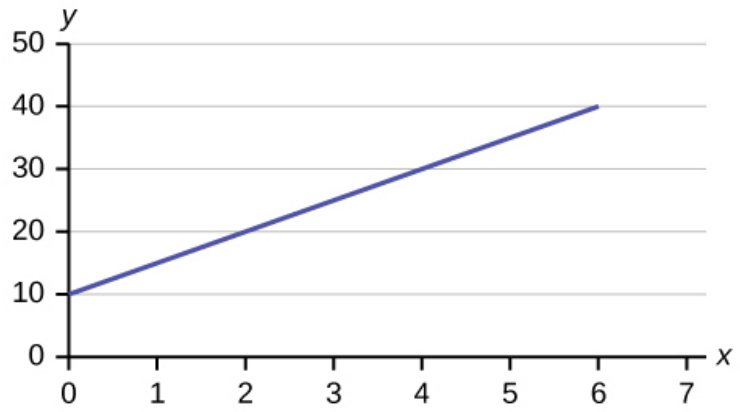
Problem:

Find the equation that expresses the total fee in terms of the number of days the payment is late.

Exercise:

Problem: Graph the equation from [link].

Solution:



Exercise:

Problem: Is the equation $y = 10 + 5x - 3x^2$ linear? Why or why not?

Exercise:

Problem: Which of the following equations are linear?

a. $y = 6x + 8$

b. $y + 7 = 3x$

c. $y - x = 8x^2$

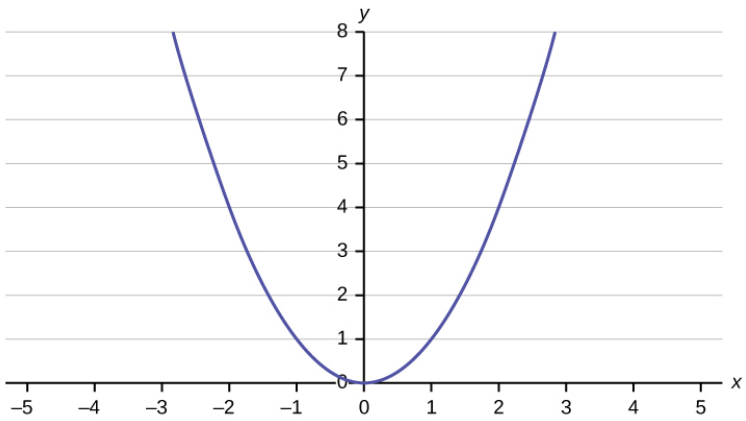
d. $4y = 8$

Solution:

$y = 6x + 8$, $4y = 8$, and $y + 7 = 3x$ are all linear equations.

Exercise:

Problem: Does the graph show a linear equation? Why or why not?



[link] contains real data for the first two decades of AIDS reporting.

Year	# AIDS cases diagnosed	# AIDS deaths
Pre-1981	91	29
1981	319	121
1982	1,170	453
1983	3,076	1,482
1984	6,240	3,466
1985	11,776	6,878
1986	19,032	11,987
1987	28,564	16,162
1988	35,447	20,868

1989	42,674	27,591
1990	48,634	31,335
1991	59,660	36,560
1992	78,530	41,055
1993	78,834	44,730
1994	71,874	49,095
1995	68,505	49,456
1996	59,347	38,510
1997	47,149	20,736
1998	38,393	19,005
1999	25,174	18,454
2000	25,522	17,347
2001	25,643	17,402
2002	26,464	16,371
Total	**802,118**	**489,093**

Adults and Adolescents only, United States

Exercise:

Problem:

Use the columns "year" and "# AIDS cases diagnosed. Why is “year” the independent variable and “# AIDS cases diagnosed.” the dependent variable (instead of the reverse)?

Solution:

The number of AIDS cases depends on the year. Therefore, year becomes the independent variable and the number of AIDS cases is the dependent variable.

Use the following information to answer the next two exercises. A specialty cleaning company charges an equipment fee and an hourly labor fee. A linear equation that expresses the total amount of the fee the company charges for each session is $y = 50 + 100x$.

Exercise:

Problem: What are the independent and dependent variables?

Exercise:

Problem:

What is the y-intercept and what is the slope? Interpret them using complete sentences.

Solution:

The y-intercept is 50 ($a = 50$). At the start of the cleaning, the company charges a one-time fee of $50 (this is when $x = 0$). The slope is 100 ($b = 100$). For each session, the company charges $100 for each hour they clean.

Use the following information to answer the next three questions. Due to

erosion, a river shoreline is losing several thousand pounds of soil each year. A linear equation that expresses the total amount of soil lost per year is $y = 12{,}000x$.

Exercise:

Problem: What are the independent and dependent variables?

Exercise:

Problem: How many pounds of soil does the shoreline lose in a year?

Solution:

12,000 pounds of soil

Exercise:

Problem: What is the *y*-intercept? Interpret its meaning.

Use the following information to answer the next two exercises. The price of a single issue of stock can fluctuate throughout the day. A linear equation that represents the price of stock for Shipment Express is $y = 15 – 1.5x$ where x is the number of hours passed in an eight-hour day of trading.

Exercise:

Problem: What are the slope and *y*-intercept? Interpret their meaning.

Solution:

The slope is –1.5 ($b = –1.5$). This means the stock is losing value at a rate of $1.50 per hour. The *y*-intercept is $15 ($a = 15$). This means the price of stock before the trading day was $15.

Exercise:

Problem:

If you owned this stock, would you want a positive or negative slope? Why?

Homework

Exercise:

Problem:

For each of the following situations, state the independent variable and the dependent variable.

a. A study is done to determine if elderly drivers are involved in more motor vehicle fatalities than other drivers. The number of fatalities per 100,000 drivers is compared to the age of drivers.
b. A study is done to determine if the weekly grocery bill changes based on the number of family members.
c. Insurance companies base life insurance premiums partially on the age of the applicant.
d. Utility bills vary according to power consumption.
e. A study is done to determine if a higher education reduces the crime rate in a population.

Solution:

a. independent variable: age; dependent variable: fatalities
b. independent variable: # of family members; dependent variable: grocery bill
c. independent variable: age of applicant; dependent variable: insurance premium
d. independent variable: power consumption; dependent variable: utility
e. independent variable: higher education (years); dependent variable: crime rates

Exercise:

Problem:

Piece-rate systems are widely debated incentive payment plans. In a recent study of loan officer effectiveness, the following piece-rate system was examined:

% of goal reached	< 80	80	100	120
Incentive	n/a	$4,000 with an additional $125 added per percentage point from 81–99%	$6,500 with an additional $125 added per percentage point from 101–119%	$9,500 with an additional $125 added per percentage point starting at 121%

If a loan officer makes 95% of his or her goal, write the linear function that applies based on the incentive plan table. In context, explain the *y*-intercept and slope.

Scatter Plots -- Linear Regression and Correlation -- MtRoyal - Version2016RevA

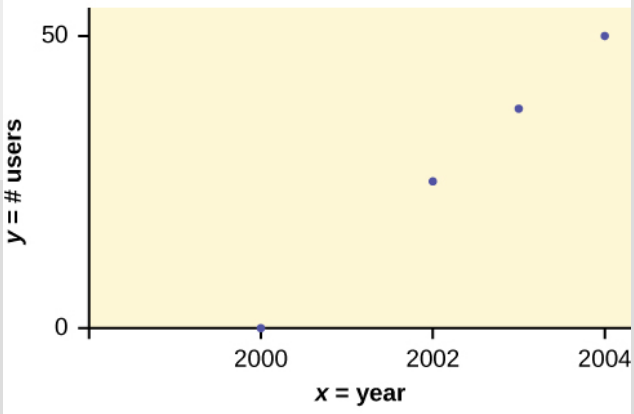
Before we take up the discussion of linear regression and correlation, we need to examine a way to display the relation between two variables x and y. The most common and easiest way is a **scatter plot**. The following example illustrates a scatter plot.

Example:
In Europe and Asia, m-commerce is popular. M-commerce users have special mobile phones that work like electronic wallets as well as provide phone and Internet services. Users can do everything from paying for parking to buying a TV set or soda from a machine to banking to checking sports scores on the Internet. For the years 2000 through 2004, was there a relationship between the year and the number of m-commerce users? Construct a scatter plot. Let x = the year and let y = the number of m-commerce users, in millions.

Table showing the number of m-commerce users (in millions) by year.

x (year)	y (# of users)
2000	0.5
2002	20.0
2003	33.0

Scatter plot showing the number of m-commerce users (in millions) by year.



x (year)	y (# of users)
2004	47.0

Note: To create a scatter plot:

1. Enter your X data into list L1 and your Y data into list L2.
2. Press 2nd STATPLOT ENTER to use Plot 1. On the input screen for PLOT 1, highlight On and press ENTER. (Make sure the other plots are OFF.)
3. For TYPE: highlight the very first icon, which is the scatter plot, and press ENTER.
4. For Xlist:, enter L1 ENTER and for Ylist: L2 ENTER.
5. For Mark: it does not matter which symbol you highlight, but the square is the easiest to see. Press ENTER.
6. Make sure there are no other equations that could be plotted. Press Y = and clear any equations out.
7. Press the ZOOM key and then the number 9 (for menu item "ZoomStat") ; the calculator will fit the window to the data. You can press WINDOW to see the scaling of the axes.

Note:
Try It
Exercise:

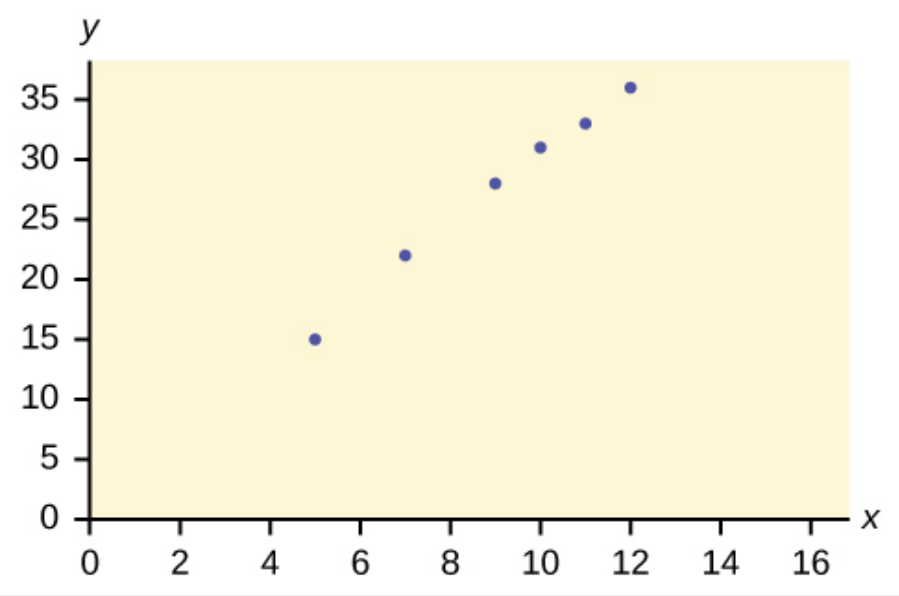
Problem:

Amelia plays basketball for her high school. She wants to improve to play at the college level. She notices that the number of points she scores in a game goes up in response to the number of hours she practices her jump shot each week. She records the following data:

***X* (hours practicing jump shot)**	***Y* (points scored in a game)**
5	15
7	22
9	28
10	31
11	33
12	36

Construct a scatter plot and state if what Amelia thinks appears to be true.

Solution:



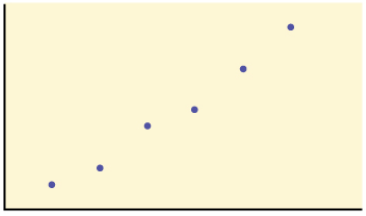
Yes, Amelia's assumption appears to be correct. The number of points Amelia scores per game goes up when she practices her jump shot more.

A scatter plot shows the **direction** of a relationship between the variables. A clear direction happens when there is either:

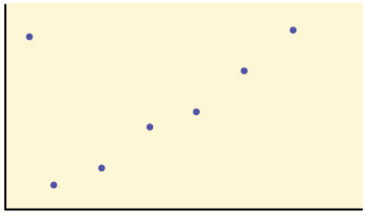
- High values of one variable occurring with high values of the other variable or low values of one variable occurring with low values of the other variable.
- High values of one variable occurring with low values of the other variable.

You can determine the **strength** of the relationship by looking at the scatter plot and seeing how close the points are to a line, a power function, an exponential function, or to some other type of function. For a linear relationship there is an exception. Consider a scatter plot where all the points fall on a horizontal line providing a "perfect fit." The horizontal line would in fact show no relationship.

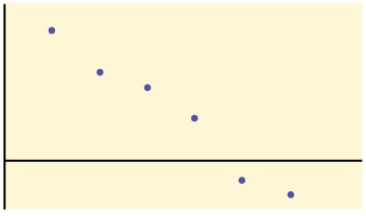
When you look at a scatterplot, you want to notice the **overall pattern** and any **deviations** from the pattern. The following scatterplot examples illustrate these concepts.



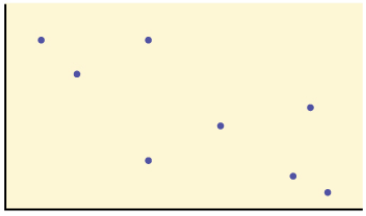
(a) Positive linear pattern (strong)



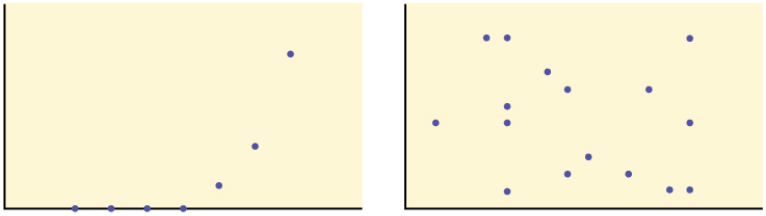
(b) Linear pattern w/ one deviation



(a) Negative linear pattern (strong)



(b) Negative linear pattern (weak)



(a) Exponential growth pattern (b) No pattern

In this chapter, we are interested in scatter plots that show a linear pattern. Linear patterns are quite common. The linear relationship is strong if the points are close to a straight line, except in the case of a horizontal line where there is no relationship. If we think that the points show a linear relationship, we would like to draw a line on the scatter plot. This line can be calculated through a process called **linear regression**. However, we only calculate a regression line if one of the variables helps to explain or predict the other variable. If *x* is the independent variable and *y* the dependent variable, then we can use a regression line to predict *y* for a given value of *x*

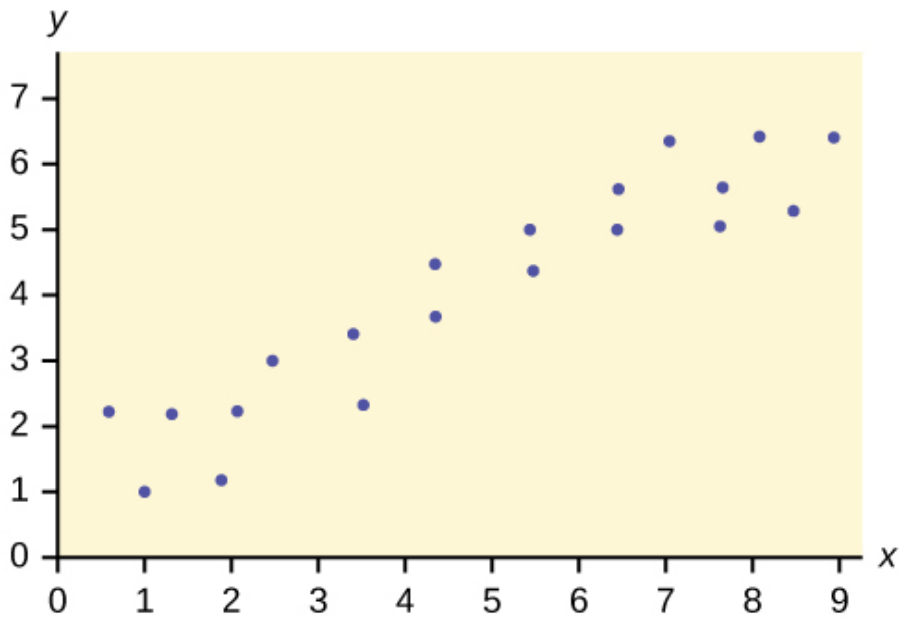
Chapter Review

Scatter plots are particularly helpful graphs when we want to see if there is a linear relationship among data points. They indicate both the direction of the relationship between the *x* variables and the *y* variables, and the strength of the relationship. We calculate the strength of the relationship between an independent variable and a dependent variable using linear regression.

Exercise:

Problem:

Does the scatter plot appear linear? Strong or weak? Positive or negative?



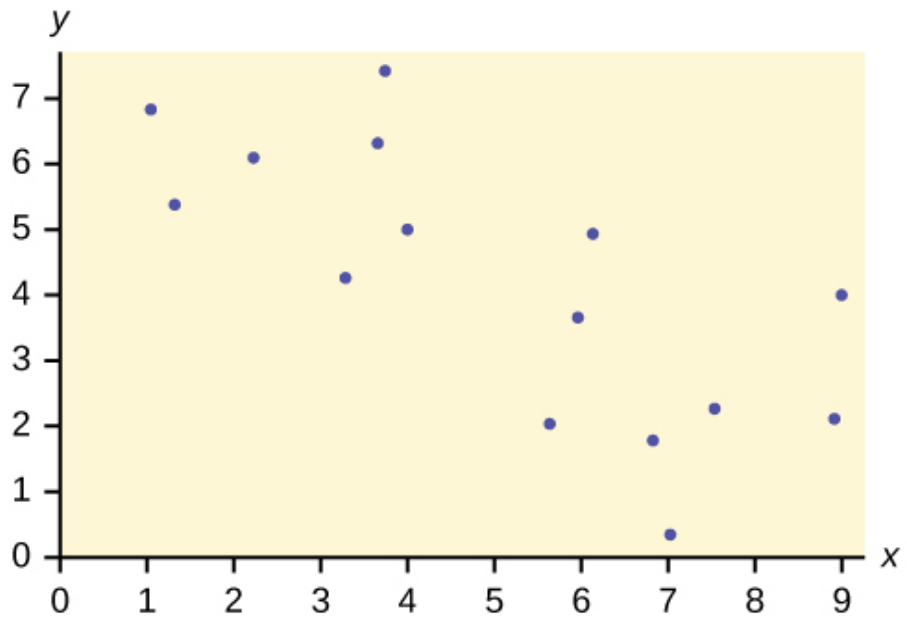
Solution:

The data appear to be linear with a strong, positive correlation.

Exercise:

Problem:

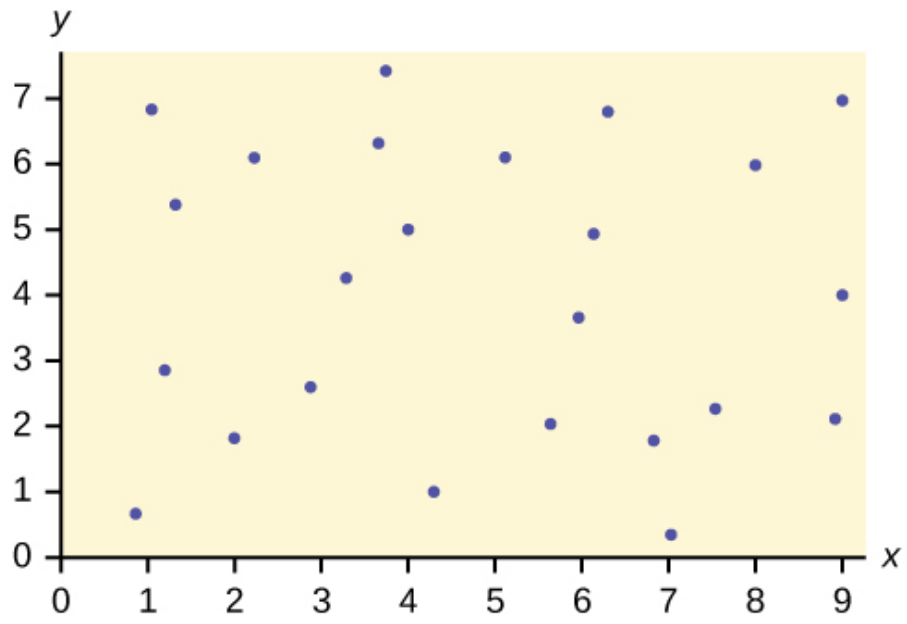
Does the scatter plot appear linear? Strong or weak? Positive or negative?



Exercise:

Problem:

Does the scatter plot appear linear? Strong or weak? Positive or negative?



Solution:

The data appear to have no correlation.

Homework

Exercise:

Problem:

The Gross Domestic Product Purchasing Power Parity is an indication of a country's currency value compared to another country. [link] shows the GDP PPP of Cuba as compared to US dollars. Construct a scatter plot of the data.

Year	Cuba's PPP	Year	Cuba's PPP
1999	1,700	2006	4,000
2000	1,700	2007	11,000
2002	2,300	2008	9,500
2003	2,900	2009	9,700
2004	3,000	2010	9,900
2005	3,500		

Solution:

Check student's solution.

Exercise:

Problem:

The following table shows the poverty rates and cell phone usage in the United States. Construct a scatter plot of the data

Year	Poverty Rate	Cellular Usage per Capita
2003	12.7	54.67
2005	12.6	74.19
2007	12	84.86
2009	12	90.82

Exercise:

Problem:

Does the higher cost of tuition translate into higher-paying jobs? The table lists the top ten colleges based on mid-career salary and the associated yearly tuition costs. Construct a scatter plot of the data.

School	Mid-Career Salary (in thousands)	Yearly Tuition

School	Mid-Career Salary (in thousands)	Yearly Tuition
Princeton	137	28,540
Harvey Mudd	135	40,133
CalTech	127	39,900
US Naval Academy	122	0
West Point	120	0
MIT	118	42,050
Lehigh University	118	43,220
NYU-Poly	117	39,565
Babson College	117	40,400
Stanford	114	54,506

Solution:

For graph: check student's solution. Note that tuition is the independent variable and salary is the dependent variable.

Exercise:

Problem:

If the level of significance is 0.05 and the *p*-value is 0.06, what conclusion can you draw?

Exercise:

Problem:

If there are 15 data points in a set of data, what is the number of degree of freedom?

Solution:

13